A Comparative Citation Analysis Study of Web-Based and Print Journal-Based Scholarly Communication —— A Look at Author Visibility in the XML Research Field

Dangzhi Zhao, Elisabeth Logan

School of Information Studies, Florida State University, Tallahassee, FL 32306-2100, USA
Dzz5355@garnet.acns.fsu.edu, ASELogan@ntu.edu.sg

Abstract:

As part of a research project that aims to identify the similarities and differences between Web-based and print journal-based scholarly communication, this paper compares author visibility revealed from citation analysis of research papers published on the Web as indexed by *ResearchIndex* with that revealed from citation analysis of journal articles as indexed by *SCI*. Results from this study demonstrate the importance and the feasibility of the use of multiple citation data sources in citation analysis studies of scholarly communication, and evidence a "two tier" future scholarly communication system.

1. Introduction

As the accelerated development of information technology, especially the rapid growth of the Web, is changing the circumstances and consequently the structures and processes of scholarly communication, there is renewed interest in the study of scholarly communication to see the types of communication that are taking place and the similarities to what we have come to expect from print based communication. Citation analysis and other bibliometric techniques have been successfully applied to the study of this new phenomenon in scholarly communication. As Zhao & Logan (2002) point out, such applications roughly fall into three categories of study. One is to apply, often with modifications, citation analysis and other bibliometric principles and techniques to study the characteristics and link structures of the Web. Examples include studies on search engines making use of hyperlink structure (Clever, 1999), and so-called "Webometrics" studies (Almind & Ingwersen, 1997; Cronin et al., 1998, Dahal, 2000; Egghe, 2000; Larson, 1996a, 1996b; Rousseau, 1997; Turnbull, 2000). The second category of studies looks at "electronic ingredients" in journal articles —— either in reference lists or in abstracts —— to see the impact of electronic publications on traditional print journal-based scholarly communication (Harter, 1992; Harter & Kim, 1996; Lu, 1999; McCain, 2000; Youngen, 1997).

A third important category of study —— citation analysis using research papers published on the Web as a data source —— has recently begun (The Open Citation Project, 2001; Goodrum et al., 2001; Zhao & Logan, 2002). Full text research papers increasingly available on the Web along with corresponding tools for searching for citations from these papers have opened up the possibility of various citation analysis studies (Cronin, 2001; Goodrum et al., 2001; Zhao & Logan, 2002). We recently reported on a study that explored this possibility and produced some interesting results. It showed that such studies are now feasible with confidence and validity levels comparable to those of traditional citation analyses based on print journals (Zhao & Logan, 2002). We also noted that our findings raised many further issues to explore, and the present study addresses some of these, building on the earlier study and seeking to further contribute to this area of inquiry.

As part of a larger research project that systematically compares scholarly communication patterns between the Web and the print world, the objectives of the present study are (1) to identify the similarities and differences in author visibility between Web-based and print journal-based scholarly communication as revealed by citation analysis; and (2) to explore possible contributing factors. The present study along with other parts of the project may contribute to the understanding of the transition of scholarly communication from print to electronic media, to the advance of citation analysis theory and methodology, and to information organization and retrieval on the Web.

2. Research questions

With a citation analysis approach, author visibility can be described in terms of how often authors have been publishing or in terms of how often their published works have been used (cited) by other scholars. Based on this consideration, the research questions to be explored in the present study are as follows.

- Are there any significant correlations between author rankings by number of publications identified from the Web and those identified from print journals in the field of XML research?
- What is the degree of correlation between author rankings by number of citations identified from the Web and those identified from print journals in the XML research field?
- What has contributed to the differences in author visibility between the Web and the print world?

The earlier study mentioned above compared author rankings between *ResearchIndex* (www.researchindex.com) and the entire *Science Citation Index* (*SCI*) database based on a visual inspection of a small set of highly visible authors. We found a considerable difference in publication patterns between these two views of the XML research field, but at the same time a strong correlation between the ten most highly cited authors. We also noted the importance of examining the characteristics of author groups with different publication and citation patterns. It is the task of the present study to

Dangzhi Zhao and Elisabeth Logan, *A comparative citation analysis study of Web-based and print journal- based scholarly communication: a look at author visibility in the XML research field*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 378-392.

examine the degree of correlation through the use of statistical approaches and more controlled data, and to explore possible contributing factors to differences in author visibility by actually examining authors' characteristics.

3. Methodology

3.1 Data collection

The ISI's *Science Citation Index* (*SCI*) and The NEC Corporation Research Institute's *ResearchIndex* are used in the present study to collect information on research papers published in print journals and on the Web, respectively. To date, the *SCI* database has been used as the data source for most of the citation analysis studies reported in the literature. *SCI* was originally designed for print journals and the majority of journals covered by *SCI* nowadays are still print-based (in print format or having a print version), although it now also selectively indexes e-journals (ISI 2000). *ResearchIndex* is a *SCI*-like tool available on the Web which automatically indexes research papers publicly available on the Web, but provides more information on cited papers than *SCI*: titles, all authors, and abstracts or full text papers for those available on the Web. More information about *ResearchIndex* can be found in Lawrence et al (1999), Goodrum et al (2001), Bar-Ilan (2001), and Zhao & Logan (2002).

Since *ResearchIndex* covers only broadly defined computer science research while *SCI* covers all sciences, three sets of source data were collected in order to control for data scope in our comparisons. They are all documents (along with their references) indexed under the term "XML" or "eXtensible Markup Language" from (1) *ResearchIndex,* (2) the entire *SCI* database, and (3) journals classified in *SCI* as representing computer science research.

Thus, the terms "XML" and "extensible Markup Language" were used to identify papers (citing papers) on XML. The actual searches were conducted on December 18, 2001. Papers that met the searching criteria were retrieved from the databases (*SCI* or *ResearchIndex*) and downloaded into a local machine. Since the existence of duplicates was found to be one of the major differences between traditional databases and the Web, paper entries retrieved from *ResearchIndex* were examined first by a Java program and then manually to remove possible duplicates. Programs were then developed in Java to convert the data formats of the retrieved paper entries to a data structure that is convenient for subsequent data analysis such as counting citations and co-citations.

Note that in the present study, the search for citing papers in ResearchIndex was limited to "Header" fields rather than searching in the full text of the documents as we had done in the previous study. The reason for this change in our data collection method was that *SCI* only goes as far as abstracts in indexing citing papers, and "Header" fields in the *ResearchIndex* database were assumed to be similar in scope. We hoped that this way of collecting data would result in more comparable data from the two data sources.

3.2 Data analysis

Three author rankings by the number of publications using fractional counts and three author rankings by the number of citations using straight counts were produced based on the three data sets respectively — the data set from *ResearchIndex*, that from the entire *SCI* database and that from a subset of *SCI* addressing computer science research.

Among the various methods of counting citations and publications, fractional counts are the most preferred and recommended by many studies (Burrell & Rousseau, 1995; Cronin & Overfelt, 1994; Egghe & Rousseau, 1990; Lindsey, 1980; van Hooydonk, 1997). The concept of fractional counts is thus chosen here for counting authors' publications. Specifically, the number of publications of each of the N authors of a publication increases by 1/N when this publication is counted. The N is supposed to be the actual number of authors of each publication. The present study however took a simplified approach in that it only took into account the first five authors rather than all authors. It was hoped that this approach would approximate strict fractional counts sufficiently as publications with more than five authors were not expected to occur too frequently based on the statistics we had (Table 1), and even if its approximation were insufficient it would still help us to see beyond the straight counts that only take into account first authors.

However, straight counts are used here for counting authors' citations because straight counts are the only citation counts supported by *SCI*. Although author rankings by fractional citation counts and by complete citation counts were obtained as well from *ResearchIndex*, results from comparisons between different citation counting methods will be reported in a separate paper.

Each of the three author rankings resulting from the different data sets is compared with the other two, and for each pair of author rankings of common authors, Pearson's r is calculated to examine the degree of correlation between the two rankings. The characteristics of three groups of authors, such as age, nationality, topic, publishing history, relationship with W3C, nature of affiliation, and collaboration orientation, are examined to identify possible contributing factors to differences in authors' visibility. The three groups are (1) authors who are highly visible both in *SCI* and in *ResearchIndex*, (2) authors who are highly visible in *SCI* but not in *ResearchIndex*, and (3) authors who are highly visible in *ResearchIndex* but not in *SCI*.

Table 1: Distribution of citing papers by number of authors				
# papers / # authors	ResearchIndex		SCI	
	#	%	#	%
0	4	1	0	0
1	83	27	75	20
2	77	25	99	26
3	78	25	86	23
4	36	12	54	14
5 or more	34	11	60	16

3.3 Results and discussion

A search on "XML" or "eXtensible Markup Language" resulted in, after removing duplicates, 312 papers using *ResearchIndex* and 374 papers with reference lists using *SCI*, 268 of which are from computer science journals. Among these papers, there are 26 common to both *ResearchIndex* and *SCI*. The papers from *ResearchIndex* made 4,578 citations, and those from *SCI* made 6,782 citations. Among the cited papers from *ResearchIndex*, 21.6% (987) are proceedings, 21.6% (991) are Web publications, 2.3% (105) are technical reports, and 2.5% (115) are from books. Among the cited papers from *SCI*, 20.6% (1399) are proceedings. We were not able to calculate the percentage of other types of documents in *SCI* due to the limited amount of information *SCI* provides about cited papers.

It can be seen that papers in both data sources in the present study are citing roughly the same percentage of proceedings and that the percentage of citing papers shared by the two data sources is very low (7% of papers in *SCI* and 8% in *ResearchIndex*). This means that in the XML research field, papers published in journals are not largely made available on the Web and papers published on the Web are not well represented in *SCI*. Since papers publicly available on the Web have been found to tend to have more impact on research (Lawrence, 2001), there appears to be a need in scholarly communication in the XML research field to promote the public availability of research papers on the Web in order to improve the efficiency of communication.

3.31. Author visibility indicated by number of publications

Three author rankings by fractional counts of publications were produced based on three data sets respectively — the data set from *ResearchIndex*, that from the entire *SCI* database and that from a subset of *SCI* addressing computer science research. They are presented in Table 2.

For reasons of convenience, only authors whose fractional publication counts are higher than one are presented in Table 2, and only the top ranked one hundred or so authors in each ranking were examined and compared. However, even though the goal was to select the top 100 authors for closer study, the number "100" was not used strictly as a cut-off but as a guideline so that authors with the same number of publications are treated the same: either all or none being selected. As a result, the number of publications used in selecting the top ranked authors was different from one ranking to another. Specifically, the criterion of "fractional publication counts greater than 0.5" was used for the ranking from the computer science journals in SCI, resulting in 101 authors, and "fractional publication counts greater than 0.8" and "fractional publication counts of at least 0.9" were used respectively for rankings from the entire SCI database and from ResearchIndex, resulting in 104 and 100 authors, respectively. These criteria were those that each resulted in a number of authors that was the closest to the goal: 100.

Among the top authors, there are 13 authors that are common to all three lists, 15 are common to the lists from *ResearchIndex* and from *SCI*, 15 are common to the lists from *ResearchIndex* and the computer science journals in *SCI*, and 77 are common to the lists from *SCI* and from the computer science journals in *SCI*. Only about 8% of the authors who actively publish are highly visible both on the Web and in print journals. This confirms at a larger scale the observation in the previous study that two very different groups of scholars are actively publishing on the Web or in journals.

Our earlier study observed that there are more authors in the Web group than in the journal group who have great impact on XML research in terms of the frequency with which they have been cited in the literature. The current data, as discussed below, show that this is true only when very highly visible authors in terms of number of publications are considered. If however authors who are not so highly visible are included, more authors in the journal group would be among the highly cited authors.

A list of highly cited authors was obtained by taking authors whose number of citations divided by the total number of citing papers in the corresponding dataset is greater than 0.018 in *ResearchIndex* or in *SCI*. This list contains 42% of the authors in *ResearchIndex* and 36% of those in *SCI* whose fractional publication counts are greater than one. However, when more authors than those whose publication counts are greater than one are concerned (for example, the top 100 or so authors in each list), the same list of highly cited authors contains fewer authors in *ResearchIndex* (19%) than in *SCI* (22%).

Table 2: Authors ranked by number of publications
(fractional counts greater than 1)

ResearchIndex		SCI		Computer science journals in SCI	
Name	**#p**	**Name**	**#p**	**Name**	**#p**
Wenfei Fan	4.07	H. S. Rzepa	4.07	A. Hunter	2
D. Fensel	2.9	P. Murray-rust	3.57	M. Rezayat	2
Dan Suciu	2.82	D. Suciu	3.2	P. T. Wood	2
Serge Abiteboul	2.68	G. V. Gkoutos	2.15	S. J. Derose	2
M. Murata	2.67	R. H. Dolin	2.05	W. Weitz	2
J. Simeon	2.45	A. Hunter	2	J. Dudeck	1.95
Angela Bonifati	2.33	A. Kristensen	2	H. S. Rzepa	1.92
Harold Boley	2.33	J. Hunter	2	S. Paraboschi	1.62
Dongwon Lee	2.25	M. Rezayat	2	P. Murray-Rust	1.58
Mark Huckvale	2	P. T. Wood	2	M. Fernandez	1.53
S. Ceri	1.9	S. J. Derose	2	H. Kim	1.5
Amarnath Gupta	1.75	W. Weitz	2	K. Canfield	1.5
Victor Vianu	1.73	J. Dudeck	1.95	N. Sundaresan	1.5
Daniela Florescu	1.7	M. F. Fernandez	1.87	E. Bertino	1.33
M. F. Fernandez	1.65	J. Simeon	1.75	F. A. Fontana	1.33
Wolfgang Emmerich	1.58	S. Ceri	1.7	G. Weikum	1.33
L. Libkin	1.5	E. Bertino	1.67	E. Damiani	1.28
Leonidas Fegaras	1.5	S. Paraboschi	1.62	L. Kerschberg	1.25
Torsten Schlieder	1.5	M. Wright	1.57	L. Rutledge	1.25
W. van der Aalst	1.5	A. Sahuguet	1.5	S. Ceri	1.2
Elena Ferrari	1.45	H. Kim	1.5	J. Simeon	1.17
John Miller	1.33	J. R. Smith	1.5	M. Shields	1.15
A. Finkelstein	1.25	K. Canfield	1.5	R. H. Dolin	1.05
B. Ludascher	1.25	N. Sundaresan	1.5		
F. Tian	1.25	C. M. Chiu	1.33		
I. Schena	1.25	F. A. Fontana	1.33		
M. Mani	1.25	G. Weikum	1.33		
Letizia Tanca	1.2	E. Damiani	1.28		
Marin Dimitrov	1.2	A. Zisman	1.25		
S. Saeyor	1.2	L. Kerschberg	1.25		
D. Kossmann	1.17	L. Rutledge	1.25		
E. Damiani	1.15	A. Y. Halevy	1.25		
S. Paraboschi	1.12	M. Shields	1.15		
Piero Fraternali	1.07				
Philip Wadler	1.03				

This indicates that among those scholars who are publishing on the Web, only very highly visible ones are likely to be recognized by the community, and that for regular scholars, publications in print journals tend to be more widely accepted. It appears that it is still of some concern among XML scholars whether work published in venues other than print journals will be recognized by the community.

If the top authors in terms of number of publications from both databases are grouped into three categories: (1) authors who actively publish both on the Web and in journals as indexed by ResearchIndex and SCI respectively, (2) authors who actively publish in journals but not on the Web, and (3) authors who actively publish on the Web but not in journals, then the same list of highly cited authors contains 47% of the authors in group 1, 19% of those in group 2, and 13% of those in group 3. Clearly, there are considerably more influential authors in the group who publish actively in both media than in those groups who only publish in one of the media. This indicates that the public availability of scholars' work on the Web can well contribute to scholars' becoming more influential, though it may not be a decisive factor.

Now the interesting question is what has contributed to this pattern of publication.

In order to find an answer to that question, data about authors' characteristics, such as age, research topics, publishing history, affiliation and collaboration orientation, were collected and examined. These data are not presented here due to limited space but can be found in Zhao (2003).

One would expect to see relatively more young scholars on the Web than in journals as it would appear easier for younger scholars to adapt to new technologies than for scholars who have a long publishing history and who therefore may not be attracted to publishing on the Web as easily. One would also expect to find more scholars on the Web than in journals who have been involved in large group collaboration as large group collaborations require open and effective communication and the Web is a perfect medium for this. It might also be expected that relatively more scholars from countries other than North America be seen on the Web as there is a well-known bias in *SCI* toward North American journals.

However, these patterns do not emerge from the data we collected in the XML research field. Actually we did not see any clear patterns about how authors' age, publishing history, nationality, affiliation or collaboration orientation have contributed to their medium preference. For example, it seems to be true with both the Web group and the journal group that there are almost as many experienced prolific scholars as young scholars with short publishing history, and that most of the highly visible scholars were professors or students when they published the articles in this study.

The only really clear factor that we were able to see is that an author's research topic is closely related to his publishing behavior. Scholars who are studying the application of computer science in general and of XML in particular tend to publish mainly in journals. For example, P. Murray-Rust is a scientist in Computational Biology (Bioinformatics, Molecular informatics) and H. S. Rzepa one in Computational Chemistry. These two top ranked scholars in *SCI* co-edited the Chemical Markup Language (CML) — a formal XML language in Chemistry. Gkoutos was Rzepa's student and was also involved in CML related research. Similarly, Dolin, who is ranked right after Gkoutos in *SCI*, is a researcher in the area of medical informatics. He has done research on XML for medical information exchange and was involved in the development of related standards. These top ranked scholars in *SCI* in terms of number of publications do not appear at all in *ResearchIndex*.

This is not difficult to understand. Scholars in an application area of a technology (e.g. computational chemistry) may have adapted to the publishing tradition within that field (e.g. chemistry) which may be different from that in the field of the technology (e.g. computer science). Although XML researchers in the computer science field are heavily publishing on the Web, scholars in the application areas of XML may not do so because they act more like, say, biologists than like computer scientists in terms of their publishing behavior.

We voiced a concern in our earlier study that some of the differences between the patterns revealed from *SCI* data on the one hand and from *ResearchIndex* data on the other, such as the more complex specialty structure derived from *SCI*, may have been due to the multidisciplinary nature of the *SCI* database. Data from the present study suggests that this should not be of great concern.

The Pearson's r for author rankings of the 113 authors shared by all three full datasets is 0.478 between *ResearchIndex* and *SCI*, 0.258 between *ResearchIndex* and the computer science journals in *SCI*, and 0.824 between *SCI* and its computer science journals. Clearly, the *ResearchIndex* data are more similar to the data from the entire *SCI* database than to that from *SCI* restricted to computer science journals. We can take that as an indication that *ResearchIndex* is really a database for computer science literature in a very broadly defined sense, and that the differences observed in our previous study between results from *ResearchIndex* and those from the entire *SCI* database cannot be explained by the multidisciplinary nature of the *SCI* database.

3.3.2. Author visibility indicated by number of citations

Three author rankings by straight citation counts were produced based on each of the three data sets — the data set from *ResearchIndex*, that from the entire *SCI* database and that from a subset of *SCI* addressing computer science research. They are presented in Table 3.

For reasons of convenience, only authors whose citation counts, divided by number of citing papers in corresponding datasets, are 0.035 or higher are presented in Table 3, and only the top ranked 100 or so authors in each ranking were examined and compared. Again, as discussed earlier, the goal was to select the top ranked 100 authors for a closer examination. However, the number "100" was not used strictly as a cut-off but as a guideline so that authors with the same number of citations are treated the same: either all or none being selected. As a result, the number of citations used in selecting the top ranked authors was different from one ranking to another. Specifically, the criterion of "6 or more citations" was used for the ranking from the computer science journals in *SCI* and that from *ResearchIndex*, resulting in 100 and 90 authors respectively, and "7 or more citations" was used for the ranking from the entire *SCI*

Table 3: Authors ranked by number of citations
(straight counts, divided by total # citing papers, 0.035 or greater)

ResearchIndex		SCI		Computer science journals in SCI	
Name	**#c**	**Name**	**#c**	**Name**	**#c**
S. Abiteboul	0.351	S. Abiteboul	0.222	S. Abiteboul	0.25
P. Buneman	0.242	T. Bray	0.206	T. Bray	0.209
A. Deutsch	0.208	A. Deutsch	0.152	P. Buneman	0.179
T. Bray	0.199	P. Buneman	0.152	A. Deutsch	0.164
J. Clark	0.186	P. Murray-Rust	0.131	J. Clark	0.127
R. Goldman	0.143	J. Clark	0.123	M. Fernandez	0.119
M. F. Fernandez	0.134	M. Fernandez	0.12	J. Robie	0.097
D. Florescu	0.115	J. Robie	0.088	Y. Papakonstantinou	0.097
Stefano Ceri	0.106	H. S. Rzepa	0.086	P. Murrayrust	0.093
J.Shanmugasundaram	0.093	Y. Papakonstantinou	0.08	R. Goldman	0.086
J. Robie	0.09	R. Goldman	0.08	S. Ceri	0.082
J. McHugh	0.087	D. Florescu	0.067	S. Cluet	0.075
Y. Papakonstantinou	0.081	R. H. Dolin	0.064	S. J. Derose	0.075
H. Thompson	0.078	S. Cluet	0.064	J. Bosak	0.071
Sophie Cluet	0.078	T. J. Berners-Lee	0.064	R. H. Dolin	0.071
S. S. Chawathe	0.071	J. Bosak	0.061	C. Goldfarb	0.067
Makoto Murata	0.068	S. Ceri	0.061	T. J. Berners-Lee	0.063
D. D. Chamberlin	0.065	C. Goldfarb	0.059	G. Wiederhold	0.06
Wenfei Fan	0.065	S. J. Derose	0.059	H. S. Rzepa	0.06
R. G. G. Cattell	0.053	D. D. Chamberlin	0.053	D. Florescu	0.056
S. DeRose	0.053	C. Friedman	0.051	P. Wadler	0.056
C. Beeri	0.05	J. Shanmugasundara	0.051	T. Milo	0.056
Tova Milo	0.05	G. V. Gkoutos	0.045	E. Maler	0.052
W. van der Aalst	0.05	T. Milo	0.045	H. Hosoya	0.052
C. Brew	0.047	A. Y. Levy	0.043	A. Bruggemannklein	0.049
H. Hosoya	0.047	G. Wiederhold	0.043	A. Hunter	0.049
O. Lassila	0.047	H. Hosoya	0.043	S. S. Chawathe	0.045
P. Wadler	0.047	J. Mchugh	0.04	C. Friedman	0.041
V. Christophides	0.047	L. Liu	0.04	D. Calvanese	0.041
E. Maler	0.043	P. Wadler	0.04	F. Neven	0.041
Angela Bonifati	0.04	S. S. Chawathe	0.04	G. Hripcsak	0.041
Jennifer Widom	0.04	D. Gardner	0.037	P. Atzeni	0.041
T. Berners-Lee	0.04	E. Maler	0.037	V. Christophides	0.037
D. Brickley	0.037	A. Bruggemannklein	0.035		
Michael Hanus	0.037	A. Hunter	0.035		
		R. GG. Cattell	0.035		
		V. Christophides	0.035		

databases resulting in 103 authors. These criteria were those that each resulted in a number of authors that was the closest to the goal: 100.

Among the top 100 or so authors from each of the three rankings, there are 49 authors that are common to all three lists, 55 are common to the list from *ResearchIndex* and that from *SCI*, 49 are common to the list from *ResearchIndex* and that from the computer science journals in *SCI*, and 84 are common to the list from *SCI* and that from its computer science journals. Through visual inspection of a small set of highly cited authors, our previous study observed a high correlation for the top 10 authors between the entire *SCI* database and *ResearchIndex*. The present study examined the correlation statistically in a much larger scale: Pearson's r's were calculated both for common authors between the three lists of top ranked 100 or so authors and for all authors that are common to all three full datasets. The Pearson's r for the 49 top ranked common authors is 0.92 between *ResearchIndex* and *SCI*, 0.91 between *ResearchIndex* and the computer science journals in *SCI* and 0.98 between *SCI* and its computer science journals. The Pearson's r's for all the 576 authors that are common to the three datasets turned out to be very similar to those for the 49 top ranked authors: 0.92 between *ResearchIndex* and *SCI*, 0.91 between *ResearchIndex* and the computer science journals in *SCI* and 0.99 between *SCI* and its computer science journals.

As seen here, the Pearson's r between author ranking from the entire *SCI* database and that from the portion of *SCI* addressing computer science research is very high. The number of top ranked authors shared by the two rankings is fairly high as well. This indicates that current XML research is still limited to computer science or that studies on the application of XML technology in different fields have been publishing in journals that are considered by *SCI* as belonging to computer science research.

The Pearson's r between author ranking resulting from *ResearchIndex* and that from *SCI* computer science data is unexpectedly lower than that between *ResearchIndex* and the *SCI* entire database although both are significant and the difference is very small. This suggests that publishing medium may have played a more important role than discipline in shaping citing authors' perceptions of cited authors' visibility in the XML research field.

Since results from the entire *SCI* database and those from computer science journals are highly correlated and the *ResearchIndex* database is more like *SCI*, which was also shown in the comparison between author rankings by number of publications discussed earlier, the following discussion will not include the dataset from the computer science journals in *SCI* anymore.

The high correlations between the author rankings indicate that when the same citation counting methods are used, a group of authors will be ranked very similarly no matter which of the two data sources is used, *ResearchIndex* or *SCI*. This confirms our interpretation of the results from our previous study as an indication that citation analysis using *ResearchIndex* as a data source is just as valid in the evaluation of scholars as citation analysis based on *SCI* data which has been widely used in the literature so far and accepted as valid in the evaluation of scholars and scholarly contributions. This also suggests that publications on the Web should not be ignored any more either as part of the literature for research or as a data source for the study of scholarly communication because they are similar to those in print journals in terms of the way they refer to earlier publications. If this can be confirmed even more strongly in the future, it is good news for bibliometric scholars who do not have access to *SCI* data, especially those in developing countries, or who investigate research areas or researcher populations under-represented in *SCI*. Now they should be able to conduct citation analysis studies using data and tools freely available on the Web and still get valid results.

However, although the correlations between author rankings are high for the common authors, common authors only account for about half of the highly cited authors (top 100 or so) from each dataset as seen from the numbers above. This indicates that the best way to evaluate scholars using citation analysis approach is to combine multiple data sources, such as *SCI* and *ResearchIndex*, so that the data sources can complement each other and the evaluation results become less biased.

Clearly there exist three groups of authors: (1) authors who are highly cited by both documents in *SCI* and those in *ResearchIndex*, (2) authors who are highly cited by documents in *SCI* but not by those in *ResearchIndex*, and (3) authors who are highly cited by documents in *ResearchIndex* but not by those in *SCI*.

Our earlier study pointed out the importance of examining the characteristics of author groups with different citation patterns. In the present study, some characteristics of the three groups of authors were examined in an attempt to identify the contributing factors to the differences in authors' visibility between *SCI* and *ResearchIndex*. Due to space constraints these data are not presented here as a table but discussed below (see Appendix H of Zhao, 2003, for complete data).

An examination of authors' characteristics suggests some interesting aspects of scholarly communication in the XML research field.

- The majority of the authors who are highly cited in both of the data sources either belong to one or more of several interrelated research groups or have been involved in World Wide Web Consortium (W3C) working groups for XML related standards or specifications.

T. Bray, J. Clark, M. F. Fernandez, J. Robie, H. Thompson, D. D. Chamberlin, M. Murata, S. DeRose, O. Lassila, E. Maler, P. Wadler, D. Brickley, V. Apparao, T. Berners-Lee, J. Bosak, S. Decker are all members of W3C working groups for XML related standards or specifications. S. Abiteboul, D. Florescu, S. Cluet, T. Milo, and V. Christophides belong to the French Project Verso group; P. Buneman, A. Deutsch, H. Hosoya, and A. Sahuguet belong to the database group at University of Pennsylvania; R. Goldman, J. McHugh, Y. Papakonstantinou, S. S. Chawathe, J. Widom, and J. Ullman belong to the database group at Stanford University (mostly the Lore project); and S. Ceri and A.

Bonifati belong to a group of Italian researchers. These groups are interrelated not only intellectually as indicated by co-citation but also socially as indicated by co-authorship.

- Foundational and historical materials and general opinion papers are highly cited in journals but not on the Web.

This can be seen quite clearly from Bosak, Goldfarb and Berners-Lee being highly cited in *SCI* but not in *ResearchIndex*.

Charles F. Goldfarb is the father of "markup languages" and the main author of the Standard Generalized Markup Language (SGML), on which the Web's HTML and XML are based. Although XML as a subset of SGML is much simpler and was designed specifically for the Web and as a result has gained wide recognition and application in the Web context, SGML is still heavily being used in publishing industry. Goldfarb's two handbooks about SGML and XML respectively are highly cited in *SCI* but not in *ResearchIndex*. Handbooks represent more mature and secondary materials and therefore are very useful in the industry. Scholars at the research front however may only refer to them as historical background as these scholars may have found the original material such as the ISO standard on SGML more convenient to use.

Tim Berners-Lee is well-known as the father of the World Wide Web and retains today an influential position as the director of the World Wide Web Consortium, while Bosak organized and led the XML working group in the development of the seminal XML specification. Both Berners-Lee and Bosak have thus laid the foundation of XML and other Web related technologies and have written influential opinion papers, such as *XML, Java, and the Future of the Web* (Bosak, 1997), *Xml and the second-generation web* (Bosak & Bray, 1999), *Weaving the Web* (Berners-Lee, et al., 1999) and *The Semantic Web* (Berners-Lee, et al., 2001).

These scholars being highly cited in *SCI* but not so much in *ResearchIndex* suggests that papers in *ResearchIndex* are perhaps more at the research front than those in *SCI* which are still referring to a considerable extent to foundational and historical materials and to opinion papers and may therefore contain more reviews and research at earlier stages.

- Authors in application areas of XML, such as Chemical Markup Language (CML) and XML for medical information exchange, are not as well represented in *ResearchIndex* as in *SCI*.

As seen from earlier discussion about authors' visibility indicated by number of publications, scholars in application areas of XML have not published often on the Web but more in journals. Examples include Murray-Rust and Rzepa who have invented CML and Friedman and Dolin who have been involved in research on XML for medical informatics. The citation patterns of scholars in these areas are shown to be very similar to the publication patterns: highly cited in *SCI* but not in *ResearchIndex*. This reveals two things clearly.

First, scholars in these areas have primarily only been cited by themselves, suggesting that these areas are relatively independent of the rest of XML research and may be quite narrow.

Second, application areas of XML are obviously not well represented on the Web. That means that citation analysis using data and tools on the Web as a data source is currently limited to certain research fields where Web publishing is well accepted by the communities, and may be biased by leaving out partly or completely specialties in which scholars have different publishing behaviors. This limitation would exist until scholarly publishing on the Web is as widely accepted and practiced as the journal.

- Authors in *the semantic Web* area and in the *Programming / processing of XML data* specialty are not as well represented in *SCI* as in *ResearchIndex*.

Many of the authors in *the semantic Web* area were ranked by number of citations much higher in *ResearchIndex* than in *SCI*. Examples include O. Lassila (24.5 in RI vs. 35.5 in SCI), D. Brickley (31 vs. 53), and D. Fensel (34 vs. 46). Some were highly cited in *ResearchIndex* but rarely or not at all in *SCI*. For example, M. Hanus and Horrocks received 12 and 7 citations in *ResearchIndex* but only 0 and 2 in *SCI* respectively. This suggests that research in *the semantic Web* area is better represented on the Web. Since this area of research is emerging and indicates the next generation of the Web, this seems to provide further evidence that research reported on the Web is perhaps more at the cutting-edge.

Among the authors who were categorized by the factor analysis routine in SPSS into the same factor representing the specialty *XML and programming*, most, including D. Megginson, N. Klarlund, R. Bourret, A. Aho, A. Schmidt, and Carl-Christian Kanne, were only highly cited on the Web. Those who were highly cited both in *ResearchIndex* and in *SCI* have been ranked much higher in *ResearchIndex* than in *SCI*. Examples include M. Murata (17 in RI vs. 35.5 in *SCI*) and D. Lee (34 vs. 46). This indicates that research in this area is also better represented on the Web than in journals.

Clearly, the discussion above about the possible bias caused by data and tools on the Web applies to *SCI* as well. In other words, citation analysis using *SCI* data as the only data source may also be biased by leaving out in part or even completely specialties in which scholars publish heavily in venues other than journals.

- Research topics are the major contributing factor to authors' different visibility in different media. Age, nationality, collaboration orientation did not seem to have made much difference.

Research on XML database design and implementation is well represented in both media. Research on XML application is better represented in journals, and that in *the Semantic Web* and *Programming / processing of XML data* areas is more visible on the Web.

It seems to be true with both media that highly cited authors are mostly from North America and Europe (especially France, Italy and United Kingdom). French and Italian researchers have been very active in research on XML database design and implementation while scholars from UK and Scotland, such as J. Clark, H. Thompson, D.

Brickley, P. Murray-Rust and H. S. Rzepa, have been actively involved in the development of many of the XML related standards or specifications.

Although there appear to be more authors among the "print only" scholars than among the "Web only" scholars who are relatively older and more experienced or less active in large group collaboration, the difference does not seem to be significant. For example, as seen from earlier discussion, both D. Suciu and W. Fan have been involved in large group collaboration but one (Fan) was only highly cited on the Web and the other (Suciu) only in journals. Another example is that young scholars can be highly cited both on the Web (e.g. Fan and Nestorov) and in journals (e.g. Dolin and Gkoutos) depending probably on the topics of their research.

4 Conclusion

Citation analysis has a long history in the study of scholarly communication and the ISI databases have until recently served as virtually the only data source for citation analysis studies. The incompleteness, bias, and limitations of this data source have been well acknowledged; nevertheless, it has remained as the basis for many studies, partly because these databases have been the only ones available for this purpose. Studies based upon ISI databases have provided valuable insight into scholarly communication patterns in many fields.

With the Web becoming a powerful communication medium, full text research papers (including reference lists) are increasingly available on the Web. Search engines and even citation indexes are emerging to help researchers make full use of these resources. It seems natural for scholarly communication researchers to be tempted to use these papers and indexes as a data source for citation analysis studies as they do not have many of the problems of the ISI databases such as the "first author only" approach to indexing cited papers with multiple authors. However, we have only seen a few studies that make use of Web data sources, and citation analysts have concerns about their use. These may include: (1) Web-publishing is not as well controlled as journal publishing and therefore might be viewed as being flawed for citation analysis. (2) Citation analysis of data from the ISI databases is considered complete enough to get a picture of scholarly communication patterns in a research field as papers indexed in these databases are considered to be "the most important" portion of the literature in the field. (3) Data and tools on the Web do not cover as many disciplines and are often not as easy to use as the ISI databases.

Findings from the present citation analysis study of XML research may help address some of these concerns.

First, the author ranking by number of citations resulting from *ResearchIndex* is highly correlated with that from *SCI*. In other words, regarding the impact of a group of scholars on research in the XML field, the collective view of citers on the Web is largely the same as that of citers in journals. Evaluation of scholars based on this view should thus be considered as equally valid, provided the discipline being studied is well published on the Web.

Second, the two groups of XML scholars who actively publish on the Web or in journals share very few publications, and are concerned with different issues. While all study XML related standards and specifications or XML database design and implementation, research on XML application is a focus only in print journals, and research on *the Semantic Web* and on *programming for, and processing of, XML data* is better represented on the Web. That means that, in order to gain a complete picture of the scholarly communication pattern in this research field, multiple data sources should be used rather than only the ISI databases or only *ResearchIndex*.

Third, although there are many advantages of using data and tools on the Web for citation analysis (Zhao & Logan, 2002), it is true that currently they do not cover as many disciplines and are not as easy to use as the ISI databases. However, these are precisely some of the aspects in scholarly communication systems that need to be improved and to which citation analysis can contribute.

For example, we can investigate how to design and implement a problem solving environment (PSE) for scholarly communication research. A PSE for scholarly communication research could put together all the computational facilities needed for studying problems in scholarly communication. It could provide a set of tools currently available on the Web, including access to *SCI* and *ResearchIndex*, and support easy integration of new data and tools, such as data filters, tools for constructing citation indexes from existing full text contents, programs for various citation and co-citation counting methods, statistical analysis tools, and visualization tools. It could also provide a graphical user interface through which scholarly communication researchers could interact with the system in the language of scholarly communication research rather than that of a certain operating system or programming language (Abrams et al., 2003; Gallopoulos, et al., 1994; Rice & Boisvert, 1996). A full discussion of this is beyond the scope of this paper, but it is clear that such a PSE would contribute both to the improvement of scholarly communication system and to the study of scholarly communication.

The differences in XML research focus in the two media as discussed above along with other findings from the present study may also shed some light on issues of scholarly communication in transition. As mentioned above, XML applications were found to be a focus only in print journals, while research into the Semantic Web and the programming for and processing of XML data was seen to be more visible on the Web. From the point of view of XML research, unlike the Semantic Web and the programming for and processing of XML data, XML applications are about relatively mature rather than cutting-edge technologies. This may be evidence of research on the Web being more at a research front than that in journals. This was also suggested by foundational and historical material being more highly cited in journals than on the Web, such as handbooks and opinion papers by Goldfarb, Bosak and Berners-Lee — some of the inventers of the Web and XML related technologies.

These results seem to be evidence of a "two-tier system" in scholarly communication that is believed by some scholars to be a probable future model of the scholarly communication system (Poultney, 1996; van Raan, 2001). In this

model, the first tier is a "free space" which represents the scholarly enterprise in "real time" and is most likely to feature free Web-based publications, while the second tier is the world of more formal publications that is most likely to continue to be dominated by journals (van Raan, 2001, p. 61).

If this system evolves, journals that currently do not accept papers published on the Web may have to change their policies, and all journals may eventually implement new procedures to reduce or eliminate the time scholars spend reformatting their research papers for journal acceptance after they have been published on the Web. This would significantly improve the efficiency of scholarly communication.

Acknowledgments

The authors wish to thank Andreas Strotmann of the Department of Computer Science, Florida State University, for his many helpful insights.

This work is supported in part by a Fellowship of the School of Computational Science and Information Technology, Florida State University.

References

Abrams, M., Allison, D., Kafura, D., Ribbens, C., Rosson, M.B., Shaffer, C., Watson L. (n.d.). PSE Research at Virginia Tech: An Overview. Retrieved January 2003, from http://research.cs.vt.edu/pse/intro.html

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: methodological approaches to "Webometrics". Journal of Documentation, 53(4): 404-426

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes — a review and analysis. Scientometrics, 50(1): 7-32

Burrell, Q., & Rousseau, R. (1995). Fractional counts for authorship attribution: a numerical study. Journal of the American Society for Information Science, 46: 97-102

Clever Project. (1999). Hypersearching the Web. Retrieved 2000, from http://www.sciam.com/1999/0699issue/0699raghavan.html

Cronin, B. (2001). Bibliometrics and beyond: some thoughts on Web-based citation analysis. Journal of Information Science, 27: 1-7

Cronin, B., & Overfelt, K. (1994). Citation-based auditing of academic performance. Journal of the American Society for Information Science, 45: 61-72

Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, Ewa. (1998). Invoked on the Web. JASIS, 49(14): 1319-1328

Dahal, TM. (2000). Cybermetics: the use and implications for Scientometrics and Bibliometrics: a study for developing science & technology information system in Nepal. Retrieved 2000, from http://www.panasia.org.sg/nepalnet/ronast/cyber.html

Egghe, L. (2000). New informetric aspects of the Internet: some reflections, many problems. Journal of Information Science, 26(5): 329-335

Egghe, L., & Rousseau, R. (1990). Introduction to Informetrics. New York: Elsevier Science Pub, 1990

Gallopoulos, E., Houstis, E., & Rice, J.R. (1994). Problem-solving environments for computational science. IEEE Computational Science & Engineering, 1: 11-23

Goodrum, A. A., McCain, K. W., Lawrence, S., & Giles, C. L. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science literature. Information Processing and Management, 37: 661-675

Harter, S. P. (1992). Psychological relevance and information science. JASIS. 43: 602-615

Harter, S. P., & Kim, H. J. (1996). Electronic Journals and Scholarly Communication: A citation and reference study. Proceedings of the Midyear Meeting of ASIS, 1996, p. 299-315

Larson, R. R. (1996a). Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. Proceedings of the 59th ASIS Annual Meeting. Baltimore, MD, Oct. 21-24, 1996, p71-78. Medford, NJ: Information Today/ASIS.

Larson, R. R. (1996b). Co-Citation analysis and the WWW. Retrieved 2000, from http://sherlock.berkeley.edu/docs/asis96/node4.html

Lawrence, S. (2001). Online or invisible? Nature, 411(6837): 521

Lawrence, S., Bollacker, K., & Giles, C. L. (1999). Digital libraries and autonomous citation indexing. IEEE Computer, 32(6): 67-71

Lindsey, D. (1980). Production and citation measures in the sociology of science: the problem of multiple authorship. Social Studies of Science, 10: 145-162

Lu, S. (1999). The transition to the virtual world in formal scholarly communication: a comparative study of the natural sciences and the social sciences. Dissertation, University of California, Los Angeles, 1999

McCain, K. W. (2000). Sharing digitized research-related information on the World Wide Web. Journal of the American Society for Information Science, 51(14):1321-1327

The Open Citation Project. (2001). Mining the social life of an eprint archive. Retrieved October 20, 2001, from http://opcit.eprints.org/tdb198/opcit/

Poultney, R. W. (1996). Front-ends are the way to go. Europhysics News, 27: 24-25

Rice, J.R., & Boisvert, R.F. (1996). From scientific software libraries to problem-solving environments. IEEE Computational Science & Engineering, 1996(Fall): 44--53.
Rousseau, R. (1997). Sitations: an exploratory study. Cybermetrics. 1(1). Retrieved October 10, 2001, from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html
Turnbull, D. (2000). Bibliometrics and the World-Wide Web. Retrieved 2000, from http://donturn.fis.utoronto.ca/research/bibweb.html
Van Hooydonk, G. (1997). Fractional counting of multiauthored publications: consequences for the impact of authors. Journal of the American Society for Information Science, 48(10): 944-945
Van Raan, A. F. J. (2001). Bibliometrics and Internet: some observations and expectations. Scientometrics, 50(1): 59-63
Youngen, G. (1997). Citation patterns of the physics preprint literature with special emphasis on the preprints available electronically. Retrieved 2000, from http://www.physics.uiuc.edu/library/preprint.html
Zhao, D. (2003). A comparative citation analysis study of Web-based and print journal-based scholarly communication in the XML research field. Doctoral dissertation, School of Information Studies, Florida State University
Zhao, D., & Logan, E. (2002). Citation analysis of scientific publications on the Web: a case study in XML research area. Scientometrics, 54(3): 449-472